W207 Final Project

Predicting Sale Price

Allison Godfrey, Jacky Ma, Surya Gutta, and Ian Anderson

Overview of our process

EDA
What does our data look like? Missing values?

Data Cleaning
How to deal with missing values? Outliers?

Feature Engineering
Which features are most relevant? How to deal with categorical and numerical features?

Create Train/ Dev Data
How to split? Use CV?

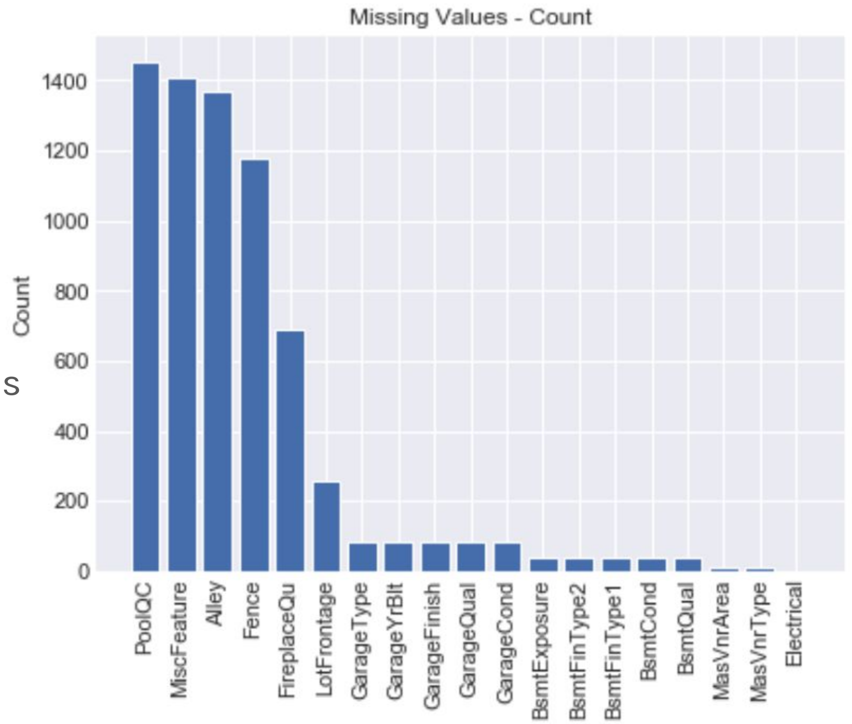
Model Setup
How does each model perform? Should we use a blended model?

Model Analysis
Which model is best? How do we measure this?

Final Submission
Base on the best performing model's RMSLE

EDA and Data Cleaning: General Approach

- Look at shape of data
  - Train (80 features, 1,460 examples)
  - Test (79 features, 1,459 examples)
- Missing Values
  - See histogram
- Look at data structure
  - Mix of categorical and quantitative features
- Use Median for Outliers
  - GrLivArea > 4000 and SalePrice < 200000
  - Total_sf > 60000 and SalePrice < 200000
  - LotFrontage > 200
  - GarageArea > 1200 and SalePrice < 300000



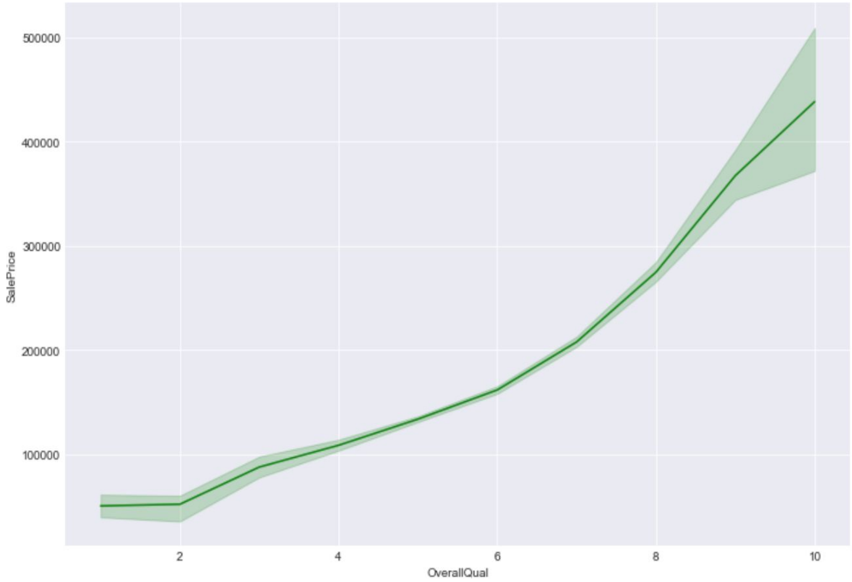
An iterative process…

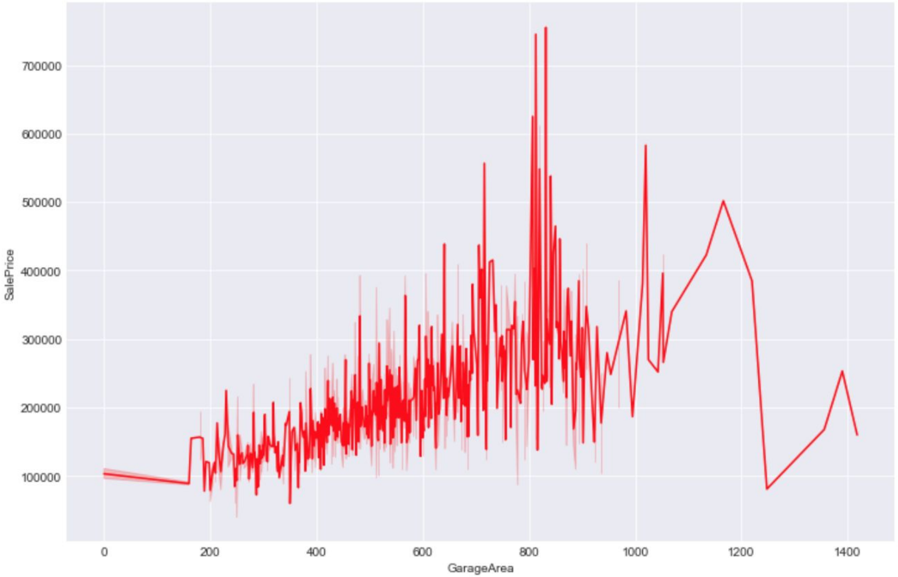
Dealing with Missing Values	*Encoding Non-Numerical Variables*	*Feature Selection*	*Selecting Models*		*Other*
Exclude all columns with > 10% missing values	LabelEncoder()	No manual feature selection	Adaboost (base estimator = LR	Random Forest	Find and label duplicates
Fill missing categorical values with NA	Hot_encode()	Take 20 most correlated features	Bayesian Ridge	Xgboost (base estimator = BR)	Use log scale for Sale Price
Fill missing categorical values with the most common category	Convert_ordinals()	Add and delete features and evaluate RMSE	Lasso Lars	Ridge	Use log scale for skewed features
Fill missing numerical values with 0		Use L1 and L2 for model feature selection	Elastic Net	Thielsen	Use 5-fold cross validation instead of 80-20 split
Fill all missing numerical values with the median value		Remove multicollinear variables	ARD	Blended Model	Tuning Hyperparameters for each model through CV
Fill select features’ missing values with median or mode		Include features with Spearman Rank > abs(0.05)			Manually sift through data points whose predictions are most off & adjust features

Features Chosen

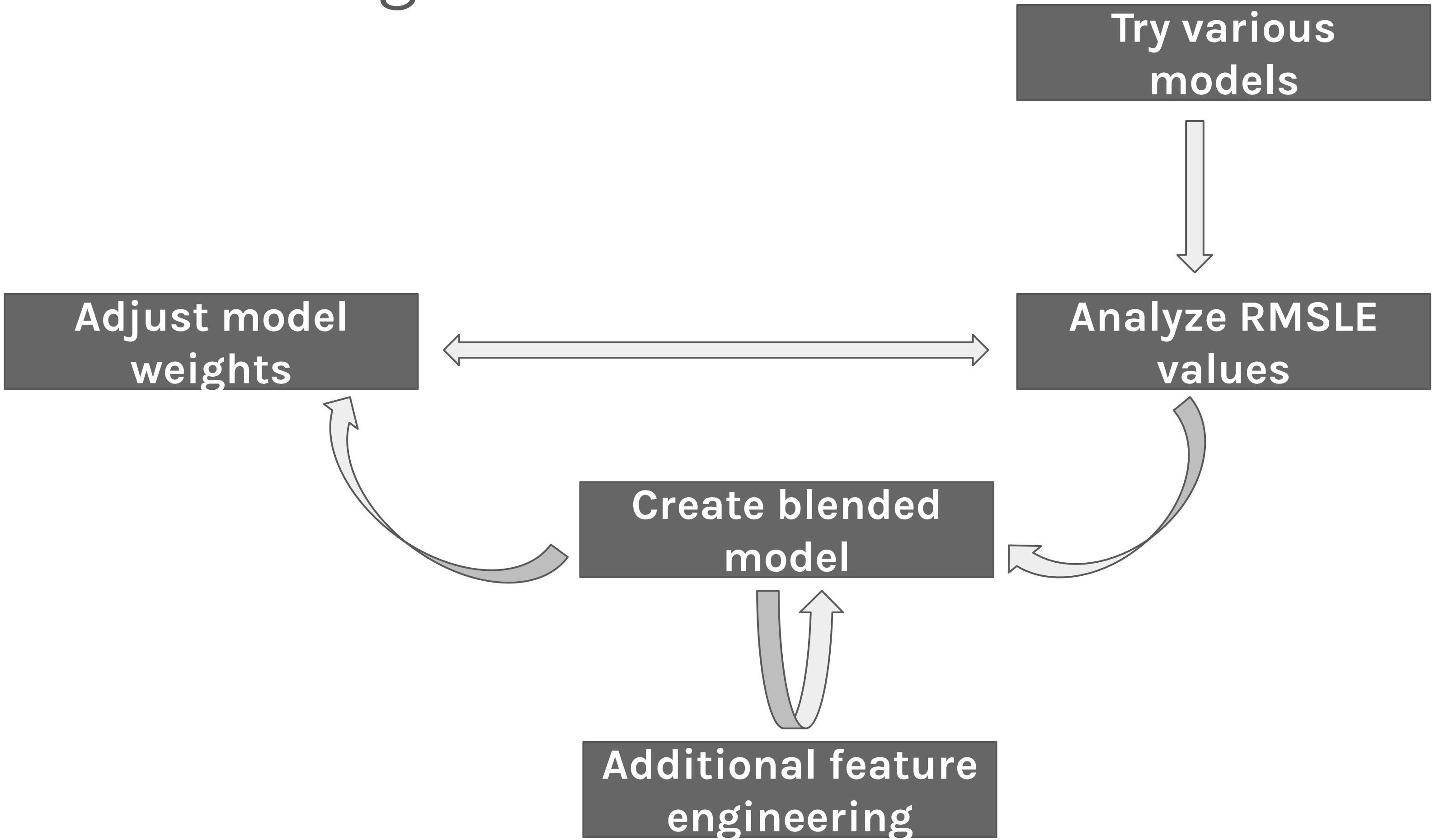
- 1stFlrSF
- 2ndFlrSF
- BldgType_X
- BsmtExposure
- BsmtFinSF1
- BsmtQual_X
- GarageType_Attc
- SaleType_X
- YrSold_X
- TotalBsmtSF
- ExterQual
- FireplaceQu_Gd
- FireplaceQu_NA
- Fireplaces
- Foundation_X
- FullBath
- Functional
- GarageArea
- GarageCars
- total_porch_sf
- GarageType_Detc
- GrLivArea
- has_fireplace
- HeatingQC
- house_age
- KitchenQual
- LotArea
- LotFrontage
- LotShape
- MasVnrType_Non
- MasVnrType_Ston
- MSSubClass_X
- MSZoning_RM
- Neighborhood_X
- OverallCond
- OverallQual
- SaleCondition_Partial
- SaleType_New
- total_baths
- MasVnrArea
- total_sf
- TotRmsAbvGrd
- WoodDeckSF
- YearBuilt
- YearRemodAdd
- GarageFinish

A closer look at specific features





Model Building Process



Converting Categorical/Ordinal Variables

```
def convert_ordinals(data):
    default_odinal = {
        'Ex': 5,  # Excellent
        'Gd': 4,  # Good
        'TA': 3,  # Average/Typical
        'Fa': 2,  # Fair
        'Po': 1,  # Poor
        'NA': 3
    }

    # LotShape: General shape of property
    lot_shape = {
        'Reg': 4,  # Regular
        'IR1': 3,  # Slightly irregular
        'IR2': 2,  # Moderately Irregular
        'IR3': 1   # Irregular
    }
```

```
def hot_encode(data):
    categorical_cols = data.select_dtypes(include=
['object'])
    return pd.get_dummies(data, columns = categoric
al_cols.columns)

test_new = hot_encode(test_new)
train_new = hot_encode(train_new)
train_new.head()
```

Using CV and tuning parameters

```
#Validation function
def cross_validation(model):
    kf = KFold(5, shuffle=True, random_state=42).get_n_splits(train_data)
    rmse= np.sqrt(abs(cross_val_score(model, train_data, train_labels, scoring="neg_mean_squared_e
rror", cv = kf)))
    return(rmse)
```

```
def parameter_tuning(model, parameters):
    clf = GridSearchCV(
        model, parameters, cv=5,scoring='neg_mean_squared_error', n_jobs = 5)

    clf.fit(train_data,train_labels)

    print(clf.best_params_)
    print(np.sqrt(-clf.best_score_))
```

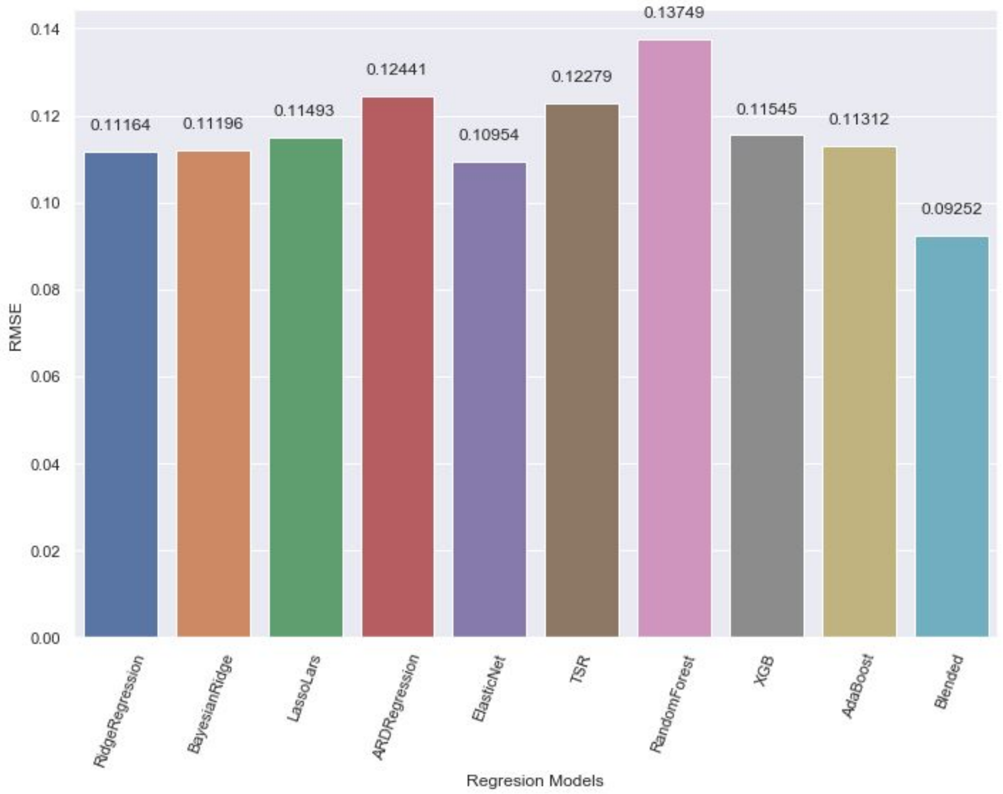
Choosing the final weights

```
regression_models = {
    'AdaBoost': {'weight': 0.125, 'model': ada_model_fit},
    'XGB': {'weight': 0.1, 'model': xgb_model_fit},
    'BayesianRidge': {'weight': 0.125, 'model': br_model_fit},
    'RidgeRegression': {'weight': 0.125, 'model': rr_model_fit},
    'TSR': {'weight': 0.125, 'model': trs_model_fit},
    'RandomForest': {'weight': 0.025, 'model': rfr_model_fit},
    'ARDRegression': {'weight': 0.125, 'model': ardr_model_fit},
    'ElasticNet': {'weight': 0.125, 'model': en_model_fit},
    'LassoLars': {'weight': 0.125, 'model': ll_model_fit},
}
```

Comparing the Models

While our RMSLE decreases with model adjustments, our score in Kaggle stays relatively constant

Are we overfitting our train data?



Our Best Kaggle Submission (with one last additional hack)

- RMSLE: 0.11725
- Place: 512

Some Future Improvements

- Further feature selection, specifically choosing the hot encoded variables to include
- More tuning of Blended Model weights
- More parameter tuning on individual models
- Evaluate PCA
- Since we know neighborhood and street, we could utilize school ratings data

Questions?